Sreyas International Journal of Scientists and Technocrats, Vol. 1(11) 2017, pp. 1-10

RESEARCH ARTICLE

Categorizing Exam Results: t-distribution

***Dr. Suresh Akella[1] Dr. Girija Akella[2]**

[1]Professor, Department of Mechanical Engineering, Sreyas Institute of Engineering & Technology
[2]Retired Scientist, Centre Research, Institute of Dry Land Agriculture (CRIDA).

Received- 9 September 2017, Revised- 10 October 2017, Accepted- 21 October 2017, Published- 30 October 2017.

ABSTRACT

A population of large output data often follows a continuous natural distribution. It is often possible to sample smaller data and to predict the population parameters. The data taker for t – distribution is samples from a population which forms a normal distribution. In this study the population of students writing examinations is divided into smaller groups of less than 30 and t-distribution is used. The data were segregated and analyzed by using MS-Excel. The t-statistic obtained can be further used for the test of Hypothesis of true population mean or similarity of means for distribution range. The error estimate 'E' is also evaluated for each category in this study. It was found that girls performed better than boys in the examination, in all categories. Also, there was no significant difference in the performance in the four sections evaluated.

Keywords: t-Distribution, Analysis of Small Data, Exam Result Analysis, Mean Comparison of Groups

1. INTRODUCTION

Large populations of variants which fit into a normal distribution occur in nature. Most of the infinite populations may be difficult to evaluate. It would be nice to divide into smaller groups of data for easier calculations. However, these groups and analysis need to be statistically valid. Thus, a population of normally distributed data is divided into smaller groups of size less than 30. These smaller groups are fit into t-distributions. Form this t-distributions statistics, size, degrees of freedom, mean, variance and standard deviation can be obtained and related to the population parameters.

t- distribution was formulated by William S. Gosset in 1908 in his evaluation of variance of mean of a distribution. As he published under the name of a student, it also became popular as Student t- test. When there is one desired output as in this case of student final results outcome of marks, a single variant t-distribution is used in this report. Sometimes, multiple outputs are required to be seen - say the yield strength, hardness elongation for a material received from different sources in a press shop. Basics of a t-distribution are given by Bhishma Rao [1]. A detailed and more theoretical analysis can be obtained in the book by Chris P. Tsokos [2]. In his book, he explained in detail both normal & t-distribution. t-distribution or Student t-distribution is a form of smaller sample data, obtained randomly from a larger population which is normally distributed. The marks of internal examinations, external examinations and final examinations for a course - Formal Languages and Automata Theory (FLAT) for II year II semester Computer Science Engineering students – were analyzed and seen that they fit well into normal distributions by Suresh Akella et. Al. [3]. Abel et al. [4] analyzed the results of European engineering examination, using a normal distribution. Similarly, the English education in different places in the United Kingdom was evaluated using the

*Corresponding author. Tel.: +919849628282
Email address: s4akella@gmail.com (S. Akella)
Double blind peer review under responsibility of Sreyas Publications
https://dx.doi.org/10.24951/sreyasijst.org/2017041001

parameters of normal distribution. European community [5] has evaluated admission process into higher education to evaluate transparent, uniform equal opportunities across Europe, using normal distribution. Golam Kibria [6] have used t-distribution for multi-variant parameters and applied it using matrix computation. Shaw and Lee [7] have formulated a bi-variant t-distribution which is a uni-variant elliptical form.

The major challenge for a statistician is finding a right selection of sample data which will be statistically close to the population. A population can have infinite elements; generally they are very large in size N. The population parameters, μ is the mean, σ is the standard deviation and σ^2 is the variance define the type of data. This data of the population can be fit into a distribution in a mathematical form to find out the probability of occurrence of an element to its expectations to the need of the evaluator. For example, in an examination of Engineering who write examinations of 4 years amount to say 400,000 for all the colleges, under the JNTUH. The data can be segregated year wise; further division to branch wise may reduce to 50 to 500. These students segregated branch wise belong to different sections, each section comprising of about 60 students. These 60 students can be grouped as male or female; and again, belonging to the type of admission they have taken. Analysis of this subgroup will give the performance of students and the management will know branch wise and group wise problems for not reaching the desired pass mark; then necessary corrective measures could be taken.

Students of a batch in a branch are subdivided into 4 sections, A, B, C and D. Further division is done as i) Boys and ii) Girls.

Next division is based on the type of admission: i) Convener quota, where the student gets admitted based on his merit in EAMCET qualifying examination; and, ii) Management category, where the student's admission is based on the Intermediate or other qualifying examinations.

Finally, the total students writing the examination in the subject FLAT become a smallest group, which has a meaning for identifying the cause of the performance. For each of these batches the distribution of frequency of occurrence is taken as a t-distribution and is analyzed. As, t distribution is applied to smaller population of < 30 size, this study can be extended to the other areas like errors in mean due to parameters of industrial products, while evaluating the relative performance.

The objectives of this study are to obtain variations in instructions, instructors, class sections, examination papers, examination dates, categories, gender, social, economic or educational backgrounds etc. of the students. Further, the corrective actions could be initiated. The mean and standard deviation comparison of the groups, from which samples are obtained, also help in the initiation of overall corrections, if found necessary, to upgrade the support lent to such categories.

2. THE t-DISTRIBUTION

Any sample data, n, should be extracted randomly from the population data of N, with or without replacement, which provides two possible distributions: one with replacement, and the other, without. It is an important assumption that population mean μ is known to apply this t-distribution to a random set of sample data which are randomly abstracted from such a population. The shape of the continuous t-distribution is also symmetric and bell shaped, the curve variation being dependent on the sample size. The mean of this distribution is 0 and the standard deviation is greater than 1, t $(\mu = 0, \sigma > 1)$ approaching 1 as the size increases to infinity when it is same as standard normal, Z $(\mu = 0, \sigma > 1)$.

While obtaining sample data, there are two methods:

1. Sampling with replacement
2. Sampling without replacement.

For the analysis in the present study, the second category is considered. Also the t-distribution is divided on the sample size n, which is $n<30$ or $n\geq30$. For the sample data the standard deviation σ is defined from the sample mean x. The t-distribution is also a symmetric distribution with mean =0 and standard deviation S >1. For the case when n tends to ∞, the standard deviation tends to 1. This is a symmetric distribution like the Standard normal distribution; but, tends to have flatter tails indicating comparatively more occurrence at the tails. The power density function or distribution frequency value of t_α is given for each risk value of α. Value of (1- α) gives confidence. Multiplied by 100 gives the % risk or confidence of evaluation.

A t-distribution graph is like the standard normal Z distribution. The t table gives the values of t_α for different values of α and the dof ν. Where, dof or DoF is the degrees of freedom of the sample data indicating to be an independent measure. It is known that $\nu = n - 1$.

3. ANALYSIS OF FINAL RESULTS OF THE COURSE FLAT FOR THE BRANCH OF COMPUTER SCIENCE AND ENGINEERING

FLAT is a course for II year II semester as under JNTUH (Jawaharlal Nehru Technological University Hyderabad). The data of the results of the students who appeared for FLAT examination final for 100 marks is shown in Figure B1. The examination is for 100 marks, a cumulative mark of the Internal examination for 25 marks and external examinations for 75 marks. The pass mark in the final assessment is 40 out of 100 marks. Akella [2] has studied the population as a normal distribution and as a standard Normal distribution as given in equation (1). In that study the distribution of Internal marks of 25, External marks of 75 and the Final total marks of 100 are shown as following Normal and also standard normal distribution. In this study the values of the population of final marks of 100 are divided into 16 categories. The first category is of sections A, B, C and D. The next Category is of Boys and Girls. Then, there is Convenor category of those who got admission through entrance examination and the Management category get admission through the marks in the Intermediate examination. These categories form 16 sample groups. It is required for t-distribution; the data are to be taken from a population belonging to a normal or standard distribution (1) which is used most often in the natural variations, manufacturing variations and was also found valid for examination results [3].

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \qquad (1)$$

When the population has a normal distribution of continuous data frequency of occurrence, for samples taken from such a data will follow t-distribution. The value of t is given as

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}} \qquad (2)$$

where t is a statistic, for different values of S and n of a sample; the numerator is an error of mean of sample, compared to the mean of the population μ. As we want this error or difference to be minimum, the numerator should be less. Similarly, the denominator is a control taking sample size n and standard deviation S, into consideration. The value of t is used for different tests of Hypothesis when sample size is less of data from a normal distribution.

4. RESULTS AND ANALYSIS

The smaller samples, 16 samples -- grouped as x1.. x16, from the population data of final marks for 100 are taken for different categories -- are taken and the required parameters of sample size n, sample mean, sample standard deviation S, statistic t and degrees of freedom ν are calculated and clubbed in tables. The population fitting to a normal distribution has a size N and population mean μ is 43.96. T-distribution is not going to use population standard deviation σ, when sample size is less. There are two varieties of sample size division:

i. When sample size is less than 30 $n < 30$
ii. When sample size $n \geq 30$

In this study the sample size is less than 30 and form 16 sample groups I = 16

Table A1 gives the Students in Section A. Boys of Convenor size n = 17, mean = 34.53; for girls n = 19 and mean = 48.79; it is clear that the girls have done better than the boys. Similarly, boys of management size n = 10 mean = 34.8; and for girls n = 9 and mean = 44.11. In both the categories of Convenor and Management quota, the girls have performed much better than the boys.

Table A2 gives the Students in Section B. Boys of Convenor size n = 17, mean = 41.76; for girls n = 19 and mean = 59.42; clearly, the girls are doing better than the boys. Similarly, Boys of management size n = 9 mean = 34.89 and for girls n = 7 and mean = 43.0. In both the categories of Convenor and Management, the results of the girls are much better than those of the boys.

Table A3 gives the Students in Section C. Boys of Convenor size n = 16, mean = 42.06; for girls n = 18 and mean = 56.83; clearly, the girls are doing better than the boys. Similarly, Boys of management size n = 8 mean = 34.88; and, for girls n = 11 and mean = 42.47. In both the categories of

Convenor and Management, the results of the girls are much better than those of the boys.

Table A4 gives the Students in Section D. Boys of Convenor size n = 16, mean = 44.25; for girls n = 17 and mean = 55.0; clearly, the girls are doing better than the boys. Similarly, Boys of management size n = 12 and mean = 34.8; and, for girls n = 9 and mean = 51.67. In both the categories of Convenor and Management, the results of the girls are much better than those of the boys.

A basic observation of the data reveals that the Boys ranking from best to least section wise of Convenor quota is DCBA; and for girls, it is BCDA. Similarly, for Management quota ranking for boys is BCAD with sections BCA having nearly the same mean value of 34.8; for girls it is CBAD. From this data it is clear that Girls performed better than Boys in all the categories. We cannot differentiate teaching quality as best or worst as sections got reversed in Convenor to Management quota. To be more specific, a test of hypothesis will enlighten better. This would be taken in a further study.

4.1 ESTIMATE OF MAXIMUM ERROR, E ANALYSIS

A statistic for error estimates the variation of population μ mean with the mean of the sample data being same as the population mean μ. This estimate gives the maximum error estimate, E. From equation (2) of statistic t we can obtain

$$t_{\alpha} \; s/\sqrt{n} = (\bar{x} - \mu) \qquad (3)$$

From equation (3) we are estimating difference or error between sample mean and population mean. To be more conservative or to obtain maximum error, we use $t_{\alpha/2}$ as $t_{\alpha/2} > t_{\alpha}$; we obtain equation (4):

$$E = t_{\alpha/2} \; s/\sqrt{n} \qquad (4)$$

The t-estimate E gives the maximum error of (μ-1) for each of the sixteen categories defined in this study. For each category of four sections, ABCD for Boys and Girls and for students in the Convenor or Management quota, the total of 16 categories is calculated. The required sample parameters are: sample size n and sample standard deviation S are taken from the calculations in Table A1 to Table A4.

The values of $t_{\alpha/2}$ are obtained from the standard tables for any value of risk factor α, for different values of dof ν. In this study α, the risk of estimate is taken as 0. 1; there is a 10% error in the estimate of this Maximum error. Otherwise, (1-α) = 0.9 there is a confidence of 90% in estimating the value of $t_{\alpha/2} = 0.05$; hence, t is obtained for different d.o.f for each of the categories of samples and listed in Table 1.

Some observations from Table 1 are: Section A has the most error variation with boys of Management quota having an estimated maximum error E, of 10.31; and, girls of Convenor category having least error of 5.44 marks from the population mean of μ = 43.96. Also, Section D has very consistent error variation between ranges of 5.41 to 6.96.

Table 1: Category of Students results of FLAT: Error Estimate E

Sec	Gender	category	$t_{0.05}$	Error, E
A	Boys	Conv.	1.746	7.39
		Mgmt.	1.833	10.31
	Girls	Conv.	1.734	5.44
		Mgmt.	1.860	7.43
B	Boys	Conv.	1.746	5.35
		Mgmt	1.860	9.08
	Girls	Conv.	1.734	4.85
		Mgmt.	1.943	8.72
C	Boys	Conv.	1.753	6.48
		Mgmt.	1.895	6.78
	Girls	Conv.	1.74	2.74
		Mgmt.	1.812	5.51
D	Boys	Conv.	1.753	6.96
		Mgmt.	1.796	5.87
	Girls	Conv.	1.746	5.59
		Mgmt.	1.86	5.41

5. CONCLUSIONS

A course, FLAT, for II year II semester CSE UG programme is found to be having normal distribution; the population of data was segregated into 16 small groups of less than 30 in number.

This data was analyzed with t-distribution. Test statistic t was calculated for different categories

and also parameters such as sample mean, size of sample dof and Standard deviation. These values are useful for test of Hypothesis to evaluate difference between means of categories or to find the confidence intervals between the two; these results will be reported in further study later.

The error analysis shows that the girls have performed better than the boys in all the categories.

There was no significant indication that a section was better than another so the teaching pedagogy is assumed to be uniform for all the sections.

ACKNOWLEDGEMENTS

Data from JNTUH & Sreyas Institute of Engineering & Technology was used for this analysis.

REFERENCES

[1] GSS Bhishma Rao, Probability & Statistics, SciTech Publications Pvt. Ltd., India.

[2] Chris P. Tsokos, Tampa, U Normal, Atlantis Studies in Probability and Statistics, Vol. 4, Series Editor & their Applications & t-distributions.

[3] Dr. Suresh Akella, Dr. P. M. Diaz, Mr. Suresh Babu, Exam Results Fit to a Normal distribution, DJ Journal of Engineering and Applied Mathematics, Vol. 4, No 1, 2018, pp. 19-25.

[4] Abel U. Osagie, Abu Mallam. Students Record Analysis and Examination Result Computation Algorithm (SRAERCA), International Journal of Technology Enhancements and Emerging Engineering Research, Vol. 2, No 8.

[5] Increasing Transparency of Access Qualifications for Higher Education in Europe 2004-3279/001-001/SO2 61-NAR, This research project report is published by UK NARIC. The project has been funded with support from the European Community.

[6] B.M. Golam Kibria, "The matrix-t distribution and its applications in predictive inference," Journal of Multivariate Analysis 97 (2006) 785 – 795.

[7] W.T. Shaw∗, K.T.A. Lee, "Bivariate Student t-distributions with variable marginal degrees of freedom and independence," Journal of Multivariate Analysis 99 (2008) 1276 – 1287.

APPENDIX A

Table A1. Population mean μ = 43.96. Samples for Section A taking FLAT Course

Sample size, n	Boys: Convenor	Boys: Management	Girls: Convenor	Girls: Management
1	9	6	21	45
2	12	32	45	51
3	56	35	55	44
4	28	52	34	63
5	16	14	52	23
6	29	42	27	47
7	15	16	34	52
8	27	60	56	29
9	16	48	31	43
10	41	43	57	
11	31		63	
12	43		69	
13	64		53	
14	46		51	
15	59		61	
16	47		60	
17	48		46	
18			65	
19			47	
Sample Size, n	17	10	19	9
Sample mean, $\bar{x}$	34.53	34.8	48.79	44.11
Sample Std. Dev. S	17.46	17.79	13.67	11.99
$t = \bar{x} - \mu) / (\frac{s}{\sqrt{n}})$	-2.23	-1.63	1.54	0.04
d.o.f. ν = n-1	16	9	18	8

Table A2. Population mean μ=43.96. Samples for Section B taking FLAT Course

Sample size (n)	Boys: Convener	Boys: Management	Girls: Convener	Girls: Management
1.	62	65	46	45
2.	24	31	54	30
3.	56	30	55	28
4.	56	14	66	44
5.	52	21	51	41
6.	33	43	58	46
7.	46	40	51	67
8.	33	40	70	
9.	41	30	81	
10.	30		52	
11.	51		55	
12.	22		44	
13.	41		70	
14.	55		63	
15.	41		74	
16.	23		79	
17.	44		46	
18.			72	
19.			42	
Sample Size, n	17	9	19	7
Sample mean, $\bar{x}$	41.76	34.89	59.42	43
Sample Std. Dev. S	12.64	14.65	12.18	11.88
$t = (\bar{x} - \mu) / (\frac{s}{\sqrt{n}})$	-0.72	-1.86	5.53	-.21
DoF ν = n-1	16	8	18	6

Table A 3. Population mean μ = 43.96. Samples for Section C taking FLAT Course

Sample size, n	Boys: Convenor	Boys: Management	Girls: Convenor	Girls: Management
1.	30	33	49	32
2.	46	33	59	52
3.	53	32	51	52
4.	21	24	47	38
5.	52	20	51	44
6.	22	46	58	27
7.	47	43	48	51
8.	35	48	56	47
9.	60		56	44
10.	65		52	64
11.	31		54	36
12.	52		64	
13.	24		60	
14.	26		63	
15.	51		72	
16.	58		61	
17.			65	
18.			57	
Sample Size, n	16	8	18	11
Sample mean, $\bar{x}$	42.06	34.88	56.83	42.27
Sample Std. Dev. S	14.80	10.12	6.67	10.08
$t = (\bar{x} - \mu) / (\frac{s}{\sqrt{n}})$	-0.51	-2.54	8.18	0.10
DoF ν = n-1	15	7	17	10

Table A 4. Population mean μ=43.96. Samples for Section D taking FLAT Course

Sample size, n	Boys: Convenor	Boys: Management	Girls: Convenor	Girls: Management
1	50	30	46	56
2	48	30	47	59
3	28	27	43	47
4	32	19	51	43
5	32	51	52	41
6	24	19	40	49
7	64	26	61	64
8	59	40	60	44
9	49	19	40	62
10	18	12	86	
11	65	40	74	
12	40	19	49	
13	48		69	
14	27		44	
15	57		45	
16	67		68	
17			60	
Sample Size, n	16	12	17	9
Sample mean, $\bar{x}$	44.25	27.67	55	51..67
Sample Std. Dev. S	15.88	11.32	13.21	8.72
$t = (\bar{x} - \mu) / (\frac{s}{\sqrt{n}})$	.07	3.27	3.44	2.65
DoF ν = n-1	15	11	16	8

APPENDIX B

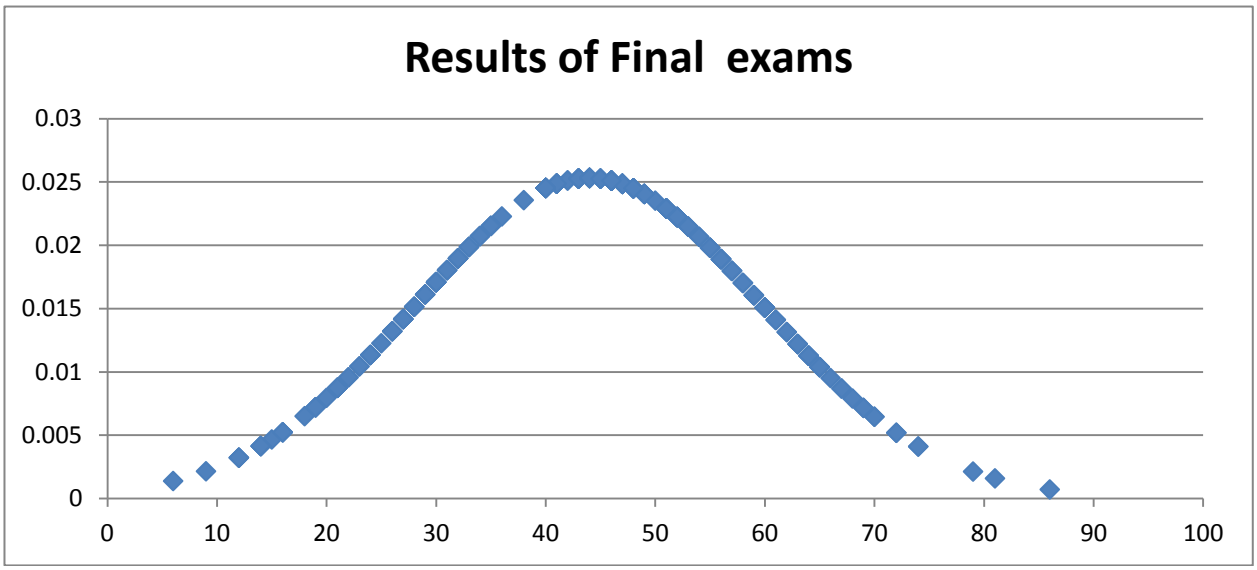

Figure B1: FLAT Course, Final Examinations results of Population